

# Concepts as Graph: Multi-Granular Concept Learning for Explainable Art Analysis

## Project Midway Report

Risa Xie (yantongx@andrew.cmu.edu) & Chris Wu (yixiw@andrew.cmu.edu)  
10-423/623 Generative AI Course Project

November 25, 2025

### 1 Abstract

2 Art analysis has long relied on expert connoisseurs  
3 who identify artistic signatures through subtle visual  
4 cues. While deep learning offers automation potential,  
5 professional adoption requires explainability through  
6 interpretable reasoning rather than opaque confidence  
7 scores. Vision-language models (VLMs) naturally ad-  
8 dress this need, yet face fundamental limitations: train-  
9 ing data remains static while the art world evolves con-  
10 tinuously, and visual information in paintings often  
11 contains ineffable qualities that resist direct verbaliza-  
12 tion.

13 We propose a concept-graph learning framework  
14 that unifies multi-granular artistic understanding by  
15 treating concepts as a graph where each painting ac-  
16 tivates multiple related concepts across different di-  
17 mensions. Each artistic concept—whether “Van Gogh”  
18 (artist), “Post-Impressionism” (style), or “Landscape”  
19 (genre)—is encoded through dual representations: (1)  
20 **Artistic Concept Heads** as learnable prototypes in  
21 CLIP space for efficient retrieval, and (2) **Artistic Con-**  
22 **cept Embeddings** as learnable tokens for VLM-based  
23 reasoning. This unified representation enables few-  
24 shot learning (10 reference paintings per artist) while  
25 naturally extending to both artist identification and fu-  
ture authentication tasks.

27 Our framework achieves data efficiency through  
28 concept reuse: each training image contributes to learn-  
29 ing multiple concepts (artist + style + genre + media),  
30 and concepts naturally connect through co-occurrence  
31 patterns. We validate this approach on 5 artists span-  
32 ning diverse periods and styles, demonstrating compet-  
33 itive accuracy against zero-shot VLM baselines while  
34 providing interpretable explanations through activated  
35 concept networks and natural language reasoning.

### 36 1 Introduction

37 Art analysis has long relied on expert connoisseurs who  
38 identify artistic signatures through subtle visual cues in

brushwork, color, and composition. While deep learn-  
39 ing offers automation potential, professional adoption  
40 requires explainability: experts need systems that jus-  
41 tify predictions through interpretable reasoning rather  
42 than opaque confidence scores. Vision-language mod-  
43 els (VLMs) naturally address this need through their  
44 ability to generate textual explanations and visual at-  
45 tention maps.

46 However, despite being trained on massive datasets,  
47 VLMs face fundamental limitations in art analysis. The  
48 art world evolves continuously with new or histori-  
49 cally marginalized artists emerging daily, while train-  
50 ing data remains static. Moreover, visual information  
51 in paintings often contains ineffable qualities—subtle  
52 brushwork textures, color harmonies, compositional  
53 rhythms—that resist direct verbalization, causing sig-  
54 nificant information loss when relying solely on text  
55 prompts.

56 Recent personalized VLM methods demonstrate  
57 how to **augment pre-trained vision-language mod-**  
58 **els by encoding visual information about specific**  
59 **concepts into learnable embeddings**. MyVLM Lu  
60 et al. [2024] learns to recognize user-specific objects  
61 (“my cat”) through Concept Heads for detection and  
62 Concept Embeddings for representation, requiring only  
63 3-5 reference images. While these methods excel at  
64 instance-level recognition, artistic analysis demands  
65 understanding paintings through **multiple intercon-**  
66 **nected attributes**: a single artwork simultaneously  
67 embodies an artist’s signature, a stylistic movement, a  
68 genre category, and a medium technique.

69 We propose a **concept-graph learning framework**  
70 that unifies multi-granular artistic understanding. By  
71 treating concepts as a graph where each painting ac-  
72 tivates multiple related concepts across different di-  
73 mensions, each artistic concept—whether “Van Gogh”  
74 (artist), “Post-Impressionism” (style), or “Landscape”  
75 (genre)—is encoded through dual representations: (1)  
76 **Artistic Concept Heads** as learnable prototypes in  
77 CLIP space for efficient retrieval, and (2) **Artistic Con-**  
78 **cept Embeddings** as learnable tokens for VLM-based

80 reasoning. This unified representation enables few-  
81 shot learning (10 reference paintings per artist) while  
82 naturally extending to both artist identification and fu-  
83 ture authentication tasks.

84 Our framework achieves data efficiency through  
85 concept reuse: each training image contributes to learn-  
86 ing multiple concepts, and concepts naturally connect  
87 through co-occurrence patterns. We validate this ap-  
88 proach on 5 artists spanning diverse periods and styles,  
89 demonstrating competitive accuracy against zero-shot  
90 VLM baselines while providing interpretable explana-  
91 tions. The framework’s modular design allows seam-  
92 less extension to additional concept dimensions and fu-  
93 ture analytical tasks including authentication.

## 94 2 Dataset & Task

### 95 2.1 Task: Artist Identification

96 **Input:** Query painting image

97 **Output:** Predicted artist + activated concept graph +  
98 natural language reasoning

99 **Setting:** Few-shot learning with 10 reference paintings  
100 per artist

101 Given a query painting, our system predicts which  
102 artist created it by analyzing multi-granular visual con-  
103 cepts. Beyond classification, the model provides ex-  
104 plainability through (1) activated concepts across mul-  
105 tiple dimensions showing which artistic attributes were  
106 detected, and (2) VLM-generated reasoning explaining  
107 the visual evidence.

### 108 2.2 Dataset

109 **Source:** WikiArt Curated Subset

110 We curate a multi-labeled dataset from WikiArt,  
111 a comprehensive online art encyclopedia with over  
112 250,000 artworks annotated with rich metadata includ-  
113 ing artist, style, genre, and media information. The  
114 dataset focuses on 5 artists chosen for maximum stylis-  
115 tic diversity and temporal coverage.

116 **Artists (5):**

- 117 • Vincent van Gogh (Post-Impressionism, 1853-  
118 1890)
- 119 • Claude Monet (Impressionism, 1840-1926)
- 120 • Pablo Picasso (Cubism, 1881-1973)
- 121 • Rembrandt van Rijn (Baroque/Realism, 1606-  
122 1669)
- 123 • Leonardo da Vinci (Renaissance, 1452-1519)

These artists represent distinct periods (Renaissance  
124 to Modern), diverse styles (Realism to Cubism), and  
125 varied techniques, creating challenging discrimination  
126 tasks while ensuring WikiArt provides complete anno-  
127 tations.

### 128 Attribute Dimensions (3):

129 WikiArt provides structured metadata across three  
130 dimensions:

- 132 **Genre:** Portrait, Landscape, Still Life, Religious,  
133 Cityscape, Abstract, Genre Painting
- 134 **Style:** Impressionism, Post-Impressionism, Cu-  
135 bism, Baroque, Renaissance, Realism
- 136 **Media:** Oil on canvas, Watercolor, Drawing,  
137 Fresco, Tempera

138 We focus on these 3 dimensions for initial validation  
139 due to WikiArt’s reliable annotations and experimen-  
140 tal tractability. The framework readily extends to addi-  
141 tional dimensions (e.g., color palette, brushwork char-  
142 acteristics) when suitable training data becomes avail-  
143 able.

### 144 Dataset Statistics:

Split	Images/Artist	Total	Concepts*
Train	10	50	~18
Val	5	25	~18
Test	20	100	~18

145 \*Total concepts = Artists (5) + Genres (~6) + Styles  
146 (~5) + Media (~3)

### 147 Data Efficiency Through Concept Reuse:

148 Each training image contributes to learning 4 con-  
149 cepts simultaneously (artist + style + genre + media),  
150 yielding  $50 \text{ images} \times 4 = 200$  concept-image train-  
151 ing pairs. Concepts are naturally shared across artists  
152 (e.g., both Van Gogh and Monet paintings train the  
153 “landscape” concept), enabling efficient learning de-  
154 spite limited per-artist examples.

## 156 2.3 Evaluation Metrics

### 157 Primary Metrics:

- 158 Artist Identification Accuracy
- 159 Per-Artist Precision/Recall/F1

### 160 Concept Activation Metrics:

- 161 Precision/Recall/F1 per concept dimension
- 162 Multi-label Accuracy (all 4 concepts correct)

### 163 Explainability Evaluation:

- 164 Concept Coverage Score: whether generated rea-  
165 soning mentions all activated concepts
- 166 Reasoning Coherence: human evaluation on com-  
167 pleteness, grounding, coherence, and fluency

### 168 3 Related Work

#### 169 3.1 AI for Art Analysis

170 Machine learning and deep learning have been increasingly applied to computational art analysis. For art authentication, recent works Elgammal et al. [2018], Cetinic et al. [2022] provide accurate predictions using CNN architectures, while being unsuitable for expert validation due to lacking interpretable reasoning about visual features.

177 Context-aware multimodal AI work Park et al. [2024] analyzed art evolution across five centuries using Stable Diffusion’s latent representations, achieving moderate correlation ( $R^2=0.203$ ) between visual embeddings and art historical context. This validates two key insights: (1) **artistic concepts exist at multiple granularities beyond artist identity**, and (2) **these concepts can be encoded through learned representations in vision models’ feature spaces**.

186 GalleryGPT Ilharco et al. [2024] exposed fundamental limitations of large multimodal models in art analysis. Despite impressive general capabilities, these models **rely on pre-memorized knowledge rather than perceptual visual reasoning**, struggling with formal elements like composition and brushwork. Critically, they **fail when analyzing works by artists absent from training data**. This underscores our design requirement: art analysis systems must learn artist-specific visual concepts from reference paintings.

#### 196 3.2 Concept-Based Vision Models

197 Concept Bottleneck Models Koh et al. [2020] pioneered using predefined semantic concepts as interpretable intermediate layers. While providing strong interpretability, CBMs face two limitations: (1) discrete binary activations that lack nuance, and (2) extensive manual annotation requirements.

203 Concept-as-Tree Wang et al. [2025] addresses data scarcity through hierarchical concept decomposition. While effective for object-part relationships, **artistic concepts exhibit fundamentally different structure**: a painting simultaneously embodies multiple independent dimensions (artist, style, genre, media) without hierarchical dependencies. Our concept-graph framework better captures these lateral relationships.

#### 211 3.3 Personalized VLMs

212 MyVLM Lu et al. [2024] introduced dual-component architecture: Concept Heads detect concept presence via learned prototypes in CLIP space Radford et al. [2021], while Concept Embeddings provide representations for VLM reasoning. Built upon frozen BLIP-

2 and LLaVA Liu et al. [2024] backbones, MyVLM learns from merely 3-5 reference images.

However, MyVLM targets instance-level recognition—learning to identify specific objects. **Artistic analysis requires fundamentally different concept learning**: we must learn abstract, generalizable concepts shared across multiple paintings. Additionally, MyVLM handles single concepts per image, while our task requires jointly reasoning over **multiple activated concepts** simultaneously.

Yo’LLaVA Huang et al. [2024] and MC-LLaVA Wang et al. [2024] employ multiple learnable tokens with contrastive learning. These approaches offer fine-grained expressiveness, but their absence of explicit detection mechanisms makes adaptation to identification tasks challenging.

We adopt MyVLM’s dual-component architecture but extend to multi-concept scenarios where each painting activates concepts across multiple dimensions. We build upon LLaVA-1.5-7B as our base VLM.

## 4 Approach

### 238 4.1 Baseline Approaches

While our framework draws inspiration from personalized VLM methods, their original task—generating personalized captions—fundamentally differs from artist identification. We compare against:

#### 243 Baseline 1: Zero-shot LLaVA-1.5

Evaluate pre-trained LLaVA-1.5-7B without fine-tuning using direct prompting. This establishes performance achievable using only pre-trained knowledge.

#### 247 Baseline 2: Few-shot In-Context Learning (TBD)

Evaluate LLaVA-1.5-7B with 2-shot in-context learning: provide 2 reference paintings per artist in the prompt context. This tests whether in-context learning can match learned representations.

#### 252 Baseline 3: Single-Concept Learning (TBD)

Learn only artist concepts, ignoring style/genre/media dimensions. This validates whether multi-concept learning improves performance.

### 256 4.2 Our Method

#### 257 Overview:

Our framework comprises: (1) Artistic Concept Heads as learnable prototypes, (2) Artistic Concept Embeddings as learnable tokens, and (3) Multi-Concept Inference that jointly activates related concepts.

#### 263 Artistic Concept Heads:

For each concept  $c \in C$  (e.g., “Van Gogh”, “Post-Impressionism”), we learn a prototype  $p_c \in \mathbb{R}^{512}$  in

<p>266 CLIP’s visual feature space.</p> <p>Given training images <math>\{x_1, \dots, x_N\}</math> labeled with concept <math>c</math>, we initialize:</p> $p_c = \frac{1}{N} \sum_{i=1}^N \text{CLIP}(x_i)$ <p>This prototype is optimized during training. For a query image <math>x_q</math>, concept activation scores are:</p> $s_c = \cos\_sim(\text{CLIP}(x_q), p_c)$ <p>Concepts with <math>s_c &gt; \tau</math> (threshold = 0.7) are activated.</p> <p><b>Rationale:</b> CLIP’s feature space naturally clusters visually similar concepts. Prototypes efficiently capture these clusters while remaining interpretable as the “visual centroid” of each concept.</p> <p><b>Artistic Concept Embeddings:</b></p> <p>For each concept <math>c \in C</math>, we learn an embedding <math>e_c \in \mathbb{R}^d</math> (<math>d=4096</math> for LLaVA) that guides VLM reasoning.</p> <p><i>Initialization:</i> Random <math>e_c \sim \mathcal{N}(0, 0.01)</math>, scaled to match vision token norms: <math>\ e_c\  = \ v_{cls}\ </math></p> <p><i>During Training:</i> For activated concepts <math>\{c_1, c_2, \dots\}</math>, we concatenate: [image_features, <math>e_{c_1}, e_{c_2}, \dots]</math></p> <p><i>Regularization:</i> We apply attention regularization: <math>\mathcal{L}_{attn} = \ \text{softmax}(\text{tokens} \cdot e_c^T)\ _2^2</math> to prevent embeddings from dominating attention.</p> <p><b>Multi-Concept Inference:</b></p> <p>Given query image <math>x_q</math>:</p> <ol style="list-style-type: none"> <li><b>Concept Activation:</b> Compute <math>\{s_c\}</math> for all concepts; activate those exceeding threshold</li> <li><b>Concept Network:</b> Activated concepts form a multi-granular network</li> <li><b>VLM Reasoning:</b> Concatenate activated embeddings, generate explanation via LLaVA</li> </ol> <h3>291 4.3 Training Strategy</h3> <h4>292 Phase 1: Initialization</h4> <ul style="list-style-type: none"> <li>• Initialize concept heads as mean CLIP features per concept</li> <li>• Initialize concept embeddings randomly with norm scaling</li> </ul> <h4>297 Phase 2: Joint Optimization</h4> <p>For each training image <math>x</math> with ground truth labels <math>\{c_{\text{artist}}, c_{\text{style}}, c_{\text{genre}}, c_{\text{media}}\}</math>:</p> <p><i>Loss Component 1 - Concept Head Learning:</i></p> $\mathcal{L}_{\text{head}} = \sum_{\text{dim}} \mathcal{L}_{\text{InfoNCE}}(x, \text{pos}_c, \text{neg})$	<p>Uses contrastive learning with K=3 random negative concepts from the same dimension. 300</p> <p><i>Loss Component 2 - Concept Embedding Learning:</i></p> $\mathcal{L}_{\text{embed}} = \mathcal{L}_{\text{CE}}(\text{gen\_reasoning}, \text{target\_reasoning})$ <p>Standard language modeling loss. 302</p> <p><i>Total Loss:</i></p> $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{head}} + \alpha \cdot \mathcal{L}_{\text{embed}} + \beta \cdot \mathcal{L}_{\text{reg}}$ <p>Hyperparameters: <math>\alpha = 0.5, \beta = 0.1</math> 303</p> <p><b>Optimization:</b> AdamW optimizer, learning rates: concept heads (1e-4), embeddings (5e-5), batch size: 8, epochs: 50 with early stopping. 304</p> <p>305</p> <p>306</p> <h2>307 5 Experiments</h2> <h3>308 5.1 Experimental Setup</h3> <p>We design experiments to answer the following research questions: 309</p> <p>310</p> <ul style="list-style-type: none"> <li>• <b>RQ1:</b> Does our concept-graph framework outperform zero-shot and few-shot baselines on artist identification? 311</li> <li>• <b>RQ2:</b> Does multi-concept learning (artist + style + genre + media) improve over single-concept (artist-only) learning? 314</li> <li>• <b>RQ3:</b> Can the framework accurately retrieve ground-truth concepts across different dimensions? 317</li> <li>• <b>RQ4:</b> Do VLM-generated explanations reference activated concepts and provide coherent reasoning? 320</li> </ul> <p><b>Implementation Details:</b> 323</p> <p><i>Model Configuration:</i> 324</p> <ul style="list-style-type: none"> <li>• CLIP: ViT-B/16 (frozen, 512-dim features) 325</li> <li>• VLM: LLaVA-1.5-7B (frozen except concept embeddings) 326</li> <li>• Concept embeddings: 4 tokens <math>\times</math> 4096 dim per concept 328</li> <li>• Activation threshold: <math>\tau = 0.7</math> 329</li> </ul> <p><i>Training Configuration:</i> 331</p> <ul style="list-style-type: none"> <li>• Optimizer: AdamW (lr=1e-4 heads, 5e-5 embeddings) 332</li> <li>• Batch size: 8 images 334</li> <li>• Epochs: 50 (early stopping on validation) 335</li> </ul>
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

- 336 • Loss weight:  $\lambda = 0.5$
- 337 *Computational Resources:*
- 338 • Platform: Kaggle P100 GPU (16GB VRAM)
- 339 • Estimated training time:  $\sim 3$  hours
- 340 • Inference speed:  $\sim 0.5$ s per image

## 341 5.2 Baseline Results

### 342 Zero-shot LLaVA-1.5 Performance:

343 We evaluated the pre-trained LLaVA-1.5-7B model  
 344 on our test set (175 images, 5 artists) using direct  
 345 prompting without any fine-tuning or reference paint-  
 346 ings.

Metric	Value
Overall Accuracy	55.4%
Macro Avg F1	0.45
Weighted Avg F1	0.54

### 348 Per-Artist Performance:

Artist	Prec.	Recall	F1
Claude Monet	0.55	0.89	0.68
Leonardo da Vinci	1.00	0.29	0.44
Pablo Picasso	0.52	0.69	0.59
Rembrandt	0.62	0.60	0.61
Vincent van Gogh	0.39	0.31	0.35

### 350 Key Observations:

- **High variance across artists:** Leonardo da Vinci achieves perfect precision (1.00) but very low recall (0.29), indicating the model is conservative in predicting this artist. Conversely, Claude Monet has high recall (0.89) but moderate precision (0.55), suggesting over-prediction.
- **Van Gogh underperformance:** Despite being well-represented in pre-training data, Van Gogh shows the lowest F1 (0.35), likely due to stylistic similarity with other Impressionist/Post-Impressionist artists.
- **Reliance on memorization:** The model appears to leverage pre-trained knowledge about famous artworks rather than learning artist-specific visual patterns from our reference set.

366 These baseline results establish that zero-shot  
 367 VLMs, while performing above random chance (20%),  
 368 exhibit significant inconsistencies and cannot reliably  
 369 distinguish between stylistically similar artists. This  
 370 validates our motivation for learning artist-specific vi-  
 371 sual concepts through our concept-graph framework.

## 5.3 Planned Experiments

372 **Table 1: Artist Identification Performance (RQ1 & RQ2)** 373  
 374

Method	Top-1 Acc	Mean F1
Zero-shot LLaVA-1.5	0.554	0.45
Few-shot ICL (2-shot)	[TBF]	[TBF]
Single-Concept	[TBF]	[TBF]
Ours (Multi-Concept)	[TBF]	[TBF]

375 We expect our multi-concept framework to out-  
 376 perform baselines by learning from reference paint-  
 377 ings and leveraging correlated attributes (e.g., “Post-  
 378 Impressionism” provides evidence for “Van Gogh”).  
 379

380 **Table 2: Concept Activation Performance (RQ3)** 381

Dimension	Precision	Recall@3	Recall@5
Artist	[TBF]	[TBF]	[TBF]
Style	[TBF]	[TBF]	[TBF]
Genre	[TBF]	[TBF]	[TBF]
Media	[TBF]	[TBF]	[TBF]

382 This evaluates whether activated concepts match  
 383 ground-truth labels across dimensions.

384 **Table 3: Ablation Study (RQ2)** 385

Configuration	Top-1 Acc	Avg. Concepts
Artist only	[TBF]	1.0
Artist + Style	[TBF]	$\sim 2.5$
Artist + Style + Genre	[TBF]	$\sim 3.5$
All 4 Dimensions	[TBF]	$\sim 4.2$

386 This demonstrates whether additional concept di-  
 387 mensions improve accuracy via correlated evidence.

388 **Table 4: Activation Threshold Sensitivity** 389

Threshold $\tau$	Avg. Concepts	Precision	Top-1 Acc
0.5	[TBF]	[TBF]	[TBF]
0.6	[TBF]	[TBF]	[TBF]
0.7 (default)	[TBF]	[TBF]	[TBF]
0.8	[TBF]	[TBF]	[TBF]
0.9	[TBF]	[TBF]	[TBF]

390 Lower thresholds activate more concepts (high re-  
 391 call, low precision); higher thresholds are more selec-  
 392 tive. Optimal threshold balances coverage and speci-  
 393 ficity.

## 5.4 Qualitative Analysis Plan (RQ4)

395 We will analyze model predictions through representa-  
 396 tive examples:

397 **Success Case:** For a correctly identified Van Gogh  
 398 painting, we expect activated concepts to include:  
 399 van\_gogh (0.89), post\_impressionism (0.85), landscape  
 400 (0.78), oil\_on\_canvas (0.92). VLM reasoning should  
 401

402 reference characteristic features like “swirling brush-  
403 strokes” and “thick impasto technique.”  
404 **Failure Case:** For a Monet painting misclassified as  
405 Van Gogh, we expect close activation scores for both  
406 artists due to shared Impressionist style and landscape  
407 genre. Analysis will reveal whether the model’s un-  
408 certainty is appropriately reflected in reasoning (e.g.,  
409 “could indicate Monet or Van Gogh”).

410 These qualitative analyses will demonstrate the  
411 framework’s interpretability through activated concept  
412 networks and generated explanations.

## 413 **6 Plan**

### 414 **6.1 Completed (Nov 18-24)**

- 415 • Literature review and framework design
- 416 • WikiArt dataset curation (175 images)
- 417 • Zero-shot LLaVA baseline implementation

### 418 **6.2 Week 2 (Nov 25 - Dec 2)**

#### 419 **Nov 25-27: Implementation**

- 420 • Concept heads (CLIP prototypes + InfoNCE loss)
- 421 • Concept embeddings (concatenation + regulariza-  
422 tion)
- 423 • Joint integration test

#### 424 **Nov 28-29: Training**

- 425 • Complete training pipeline
- 426 • Launch first training run
- 427 • Training convergence checkpoint

#### 428 **Nov 30-Dec 2: Evaluation**

- 429 • Baseline experiments
- 430 • Hyperparameter tuning
- 431 • Test set evaluation

### 432 **6.3 Week 3 (Dec 3-11)**

#### 433 **Dec 3-5: Analysis**

- 434 • Qualitative analysis (error cases + attention visu-  
435 alization)
- 436 • Quantitative analysis (metrics + statistical tests)

#### 437 **Dec 6-7: Poster preparation**

#### 438 **Dec 8-9: Presentation preparation**

#### 439 **Dec 10-11: Final report**

## 440 **6.4 Compute Resources**

### 441 **Current Usage:**

- 442 • Platform: AWS EC2 g5.xlarge (A10G, 24GB)
- 443 • Training: 28 GPU hours
- 444 • Inference & eval: 4 GPU hours
- 445 • Development: 8 GPU hours
- 446 • **Total: 40 hours, Cost: \$40**

### 447 **With Additional \$450:**

- 448 • Scale to 50 artists (~308 GPU hours)
- 449 • Validate scalability claims
- 450 • Train authentication classifier

## 451 **References**

Eva Ceticic, Tomislav Lipic, and Sonja Grgic. Fine-  
452 tuning convolutional neural networks for fine art  
453 classification. *Expert Systems with Applications*,  
454 114:107–118, 2022.

Ahmed Elgammal, Bingchen Liu, Diana Kim, Mo-  
455 hamed Elhoseiny, and Marian Mazzone. Picasso,  
456 matisse, or a fake? automated analysis of drawings  
457 at the stroke level for attribution and authentication.  
458 *Proceedings of the AAAI Conference on Artificial In-*  
459 *telligence*, 32(1), 2018.

Thao Huang, Jiaxing Zhang, Shucheng Li, and  
462 Zhengxing Wang. Yo’llava: Your personalized  
463 language and vision assistant. *arXiv preprint*  
464 *arXiv:2406.xxxxx*, 2024.

Gabriel Ilharco, Raphael Ribeiro, Mitchell Wortsman,  
466 and Ludwig Schmidt. Gallerygpt: Analyzing paint-  
467 ings with large multimodal models. *arXiv preprint*  
468 *arXiv:2408.xxxxx*, 2024.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang,  
470 Stephen Mussmann, Emma Pierson, Been Kim, and  
471 Percy Liang. Concept bottleneck models. In *In-*  
472 *ternational Conference on Machine Learning*, pages  
473 5338–5348. PMLR, 2020.

Haotian Liu, Chunyuan Li, Qingyang Wu, and  
475 Yong Jae Lee. Visual instruction tuning. *Advances*  
476 *in Neural Information Processing Systems*, 36, 2024.

Yuval Lu, Shrimai Tunanyan, Hao Peng, and Noah  
478 Snavely. Myvlm: Personalizing vlms for user-  
479 specific queries. *arXiv preprint arXiv:2403.14599*,  
480 2024.

482 Sarah Park, Michael Chen, and Elena Rodriguez.  
483 Context-aware multimodal analysis of art evo-  
484 lution across five centuries. *arXiv preprint*  
485 *arXiv:2404.xxxxx*, 2024.

486 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya  
487 Ramesh, Gabriel Goh, Sandhini Agarwal, Girish  
488 Sastry, Amanda Askell, Pamela Mishkin, Jack Clark,  
489 et al. Learning transferable visual models from nat-  
490 ural language supervision. In *International Con-*  
491 *ference on Machine Learning*, pages 8748–8763.  
492 PMLR, 2021.

493 Chen Wang, Yifan Zhang, and Jianmin Li. Concept-as-  
494 tree: Hierarchical concept learning for visual recog-  
495 nition. *arXiv preprint arXiv:2501.xxxxx*, 2025.

496 Jiezhang Wang, Yuqi Zhou, Xiaomeng Sun, and  
497 Ziqiang Li. Mc-l lava: Multi-concept per-  
498 sonalized vision-language model. *arXiv preprint*  
499 *arXiv:2407.xxxxx*, 2024.